Marathon Performance Prediction of Amateur Runners based on Training Session Data

Daniel Ruiz-Mayo, Estrella Pulido, and Gonzalo Martínez-Muñoz

Escuela Politécnica Superior,
Universidad Autónoma de Madrid,
Madrid, Spain

Abstract. In this article we analyze the use of machine learning algorithms to predict the marathon performance of amateur runners based on their training regimes. We gathered data about marathon marks of amateur athletes and about their workouts in the six week period prior to the race. The data included information of athletes running the 2015 Boston and London marathons. Every workout includes information about distance, time, net elevation gain and pace. The training regime of each runner was summarized in a standardized vector in order to be able to feed it to machine learning algorithms. To perform this standardization we propose to split the workouts into one week, two week and six week periods and to extract different statistics of the trainings within those periods. The proposed statistics include the total number of kilometers run in the period, the fastest pace or the longest distance in a single workout. An empirical analysis under several conditions was carried out using different sets of runners and races. The results show that, for the studied datasets and proposed set of attributes, bagging provides rather good estimates of the performance of amateur athletes in a challenging and unpredictable race, such as the marathon.

1 Introduction

Over the last decade running has become a widespread physical activity [10, 11, 9, 8]. There are about 50 million Europeans that run as a way to be fit [8]. In Spain, the number of runners during the decade 2000 to 2010 experienced an increment of 50% reaching 5.6% of the population (from ages 15 to 74) that practice running. This increment in the number of runners comes together with a step growth in the number of runners that prepare for a marathon. The number of marathons and marathon runners worldwide has experienced a sharp increase in the last 15 years from nearly 1000 marathons (and about 400,000 finishers) in 2000 to almost 4000 marathons (and about 1,600,000 finishers) in 2013 [9]. This is a 400% increase in the number of marathon events and finishers worldwide in only 13 years. In the US during the same period, the increment of marathon finishers was more moderate (about 50% increase from 2000 to 2013) [11]. This is probably because in the US, the number of runners is arriving already to a saturation point.

In this study we will focus on the analysis of marathon mark prediction. Running a marathon involves a fairly large period of preparation of several months. Despite of this long period of training, this challenging race is quite unpredictable for amateur runners. Many factors can affect the performance during the race, such as cramps, starting at a pace too fast for his/her preparation, dehydration, etc. Having a good estimate of what their performance could be, could help marathon amateur runners to better face the race. There exist some performance predictors, such as [1], that are based on relative performance tables across events [5, 6]. The idea is that equivalent performances in different races would be assigned the same number of points. In [5], the point system has, as a reference, the estimations of maximum speeds that could be reached in different events (Portuguese tables). Instead of using a single reference value for each event (i.e. the maximum expected speeds) for comparing events, Godsey proposes to use the set of top performances [6]. To get a prediction of your mark using these tables, one has to run another event (e.g. a half-marathon), get the points assigned to one's performance for that event and finally calculate the marathon mark assigned to the same number of points [1]. This method assumes that the performances of a single athlete in different distances are equivalent, which is often not the case. In addition, this method does not take into account how fit the athletes are prior to the race.

In this paper, we propose to use machine learning models to estimate the performance of amateur runners based on their training regimes during the weeks prior to the marathon. To train the machine learning models, we propose to use attributes that encode the information about their training sessions such as the distance covered, the net elevation gain or the pace.

The articles is organized as follows. In Section 2, the details of the proposed methodology are described. Section 3 shows the results of the empirical evaluation of the proposed method. Finally, the conclusions of this investigation are summarized in Section 4.

2 Proposed methodology

This paper proposes an experimental study for predicting the marks obtained by amateur marathon runners based on data collected about their training sessions. Additional more specific goals are the following:

- Propose a set of attributes to adequately characterize each runner training regime.
- To apply machine learning algorithms to predict their marks in a marathon based on the proposed characterization of their training sessions.
- To compare the performance of different machine learning algorithms under different sets of attributes and experimental conditions.

For these analysis we used the training data from hundreds of athletes running the 2015 London and Boston marathons. This data, along with the information about the finishing time in the marathon, has been extracted from the public profiles of runners in a sports social network. Different techniques have been used for the extraction process such as crawlers, headless browsers and bash scripting. The information gathered contains all the training sessions posted during the 6 week period prior to the race. Each training session includes the date, distance, time, pace and elevation gain. Table 1 shows the information of the extracted dataset including the number of runners, number of workouts and average workouts per athlete for the Boston and London marathons.

Table 1. Unfiltered dataset information (more detail in the text)

	# of runners	# of workouts	workout/runner
London	1766	52626	29,8
Boston	1620	52304	32.3

It is reasonable to think that people preparing a marathon should at least train three days per week to be in a reasonable good shape for the race. Actually, the number of recommended workouts per week to prepare for a marathon is above that number in most plans (see for instance [4]). However, after the analysis of the main characteristics and statistics of the marathon data we observed that there were many athletes with a low number of workouts in some weeks of the 6 week period. This reduced number of trainings in some weeks might be due to the fact that they do not post all their workouts or that they post them in a different latter date. In order to have a dataset closer to what we should expect from a person preparing a marathon, we have removed those athletes with a low number of workouts. In particular, we have removed the athletes who have any week with less than three workouts except for the week prior to the marathon. Furthermore, the original dataset has some workouts with extremely long distances (over 50km) and high paces (faster than 2:00 min/km) that might correspond to bike workouts uploaded as running workouts. These workouts have been also eliminated. Table 2 shows the information of the resulting dataset. As it can be observed the number of runners has been reduced by half for Boston and by 8 for London after filtering. For those runners that remain the average number of workouts in the six week period is increased to approximately 40 workouts/runner.

Table 2. Filtered dataset information (more details in the text)

	# of runners	# of workouts	workout/runner
London	208	8541	41.1
Boston	792	27068	34.2

Once we have the information about the workouts for the runners and their marathon marks, the next step is to feed the machine learning algorithms with this information. For this purpose it is necessary to transform the information about the runners workouts into a homogeneous vector of attributes. This vector should contain a fixed set of attributes for all athletes independently of the number of workouts each of them had performed. Our approach is to divide the whole six week training period prior to the marathon into shorter periods and to extract attributes from those periods. The attributes will be statistics about the training sessions in those time periods. Several time periods have been defined that give rise to three different sets of attributes. The considered time periods treat the training sessions per week, every two weeks, and as a whole. For each time period the following attributes are extracted:

- Distance of the training session with the highest distance covered.
- Time of the training session with the highest distance covered.
- Pace of the training session with the highest distance covered.
- Net elevation gain of the training session with the highest distance covered.
- Distance of the training session with the highest pace.
- Time of the training session with the highest pace.
- Pace of the training session with the highest pace.
- Net elevation gain of the training session with the highest pace.
- Total number of kilometers covered in the training sessions for the given period
- Total time of the training sessions.
- Total net elevation gain for the training sessions.

With this attributes we will build three attribute sets: one containing the six one week periods, a second one containing 3 two week periods and a combined set of attributes. The combined set of attributes contains all attributes from the weekly and two-week sets of attributes and also from the period containing all the training sessions. In addition, the age and sex of the runner was also included as attributes. In summary, there will be $11 \times 6 + 2$ attributes for the weekly attribute set, $11 \times 3 + 2$ for the two-week attribute set and $11 \times (6 + 3 + 1) + 2$ for the combined attribute set.

3 Experiments

In order to analyze the validity of the selected attributes for the estimation of the marathon performance we have carried out a series of experiments. For all experiments, the validation is done using 10-fold cross-validation. The regression methods used are linear regression and bagging [3]. All experiments were performed using Weka [7] and Knime [2]. Linear regression was applied by using the default parameters (i.e. M5 attribute selection method, eliminate colinear attributes set to true and a value for ridge of $1E - 8$). For bagging 100 regression trees are combined. As a baseline, ZeroR regressor was also applied. This method predicts, for all the instances, the mean of the objective variable in the training set.

For the first experiment, we have used the original data containing all available runners independently of the number of training sessions that they performed. Three sets of attributes were used as described in the previous section: weekly, two-week and combined. The combined statistics included the weekly and the two-week period attributes and also the statistics for the whole 6-week period of available data. We used two datasets that contain the statistics of the training sessions of runners that run in the London and the Boston marathon. In order to have more data available for training, we also merged both datasets.

The results for this experiment are shown in Table 3. In this table the average mean absolute error (in seconds) is shown for the London and Boston marathons individually and for both. Each column corresponds to the results of the different tested algorithms (ZeroR, Linear regression and Bagging) for each set of attributes: weekly (shown as week), two-week (as 2-week) and combined (as comb).

Table 3. Mean absolute error (in seconds) for unfiltered data

	ZeroR	Linear Regression			Bagging		
	-	week	2-week	comb	week	2-week	comb
London	1260.1	867.7	873.0	855.6	717.3	699.1	681.2
Boston	1449.4	1161.2	1130.1	886.4	866.3	825.5	712.9
Both	1335.6	941.0	928.0	863.1	766.6	731.6	707.8

Several interesting results can be observed in Table 3. Among the tested methods, bagging achieves the best performance since it obtains the best mean absolute error for each different set of attributes and marathon. The performance of linear regression is between 22% and 37% worst than the performance of bagging and the differences are over 150 seconds in all cases. This suggest that regressors that are more complex than linear are needed to obtain better estimates for marathon performance. Further analysis should be carried out to analyze the importance of the different attributes in the performance prediction.

In addition, the improvement with respect to the baseline, ZeroR, is noticeable. Observe that ZeroR does not use the attributes to predict the output variable. Its output is simply the average output in the training set. Hence, this improvement with respect to ZeroR indicates that the selected attributes contain relevant information for the prediction of marathon performance. Finally, by studying in detail the differences between the selected sets of attributes, we can observe that the combined set of attributes is the one that produces the best results for each method and marathon. This seems reasonable since the combined set of attributes contains all the attributes of the weekly and two-week sets. Given the results shown in Table 3, only the combined set of attributes will henceforth be used.

The results shown in Table 3 used the information from all runners independently of the number of training sessions available for each of them. Running a marathon involves an important amount of training [4]. For this reason, we have repeated the experiments using only the information about runners for which we have information of at least three training seassons per week. This condition was set for the weeks prior to the race except for the last one where no minimum training sessions was set.

Table 4 shows the mean absolute error (in seconds) for the filtered data. The results shown correspond to the combined set of attributes. From this table we can observe an improvement with respect to the unfiltered data (shown in Table 3, column "comb"). Bagging using both datasets achieves the best results with an average mean error of 555.0 seconds ($\approx$ 9 minutes).

Table 4. Mean absolute error (in seconds) for filtered data

	ZeroR	Linear Regression	Bagging
London	920.1	762.1	564.3
Boston	1019.3	712.1	582.9
Both	1054.6	658.0	555.0
Injure-free	900.6	530.4	436.3

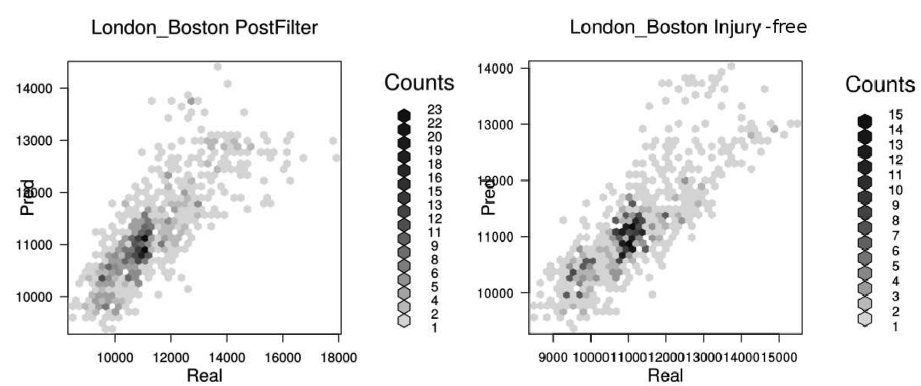

Fig. 1. Hexbin plots of the predictions vs. the real values for London-Boston filtered data (left plot) and *injury-free* data (right plot)

Finally, we have carried out an experiment aimed at estimating the performance of runners who did not run into trouble during the race. The idea is to take into account the many factors that can impair the performance in a marathon and remove the runners who were affected by them. Among these factors we can consider deviating from the right pace, cramping, or dehydrating, among others. To simulate this situation, we have removed those athletes with a performance 900 seconds below the one estimated with the bagging method trained on the filtered data. Another way of detecting these athletes would be by using the performance drop in the second half of the marathon with respect to the first half. However, the split times were not available. The results for this subset of runners are shown in the last row of Table 4 under *injure-free*. The performance of this last experiment is of 436.3 second for bagging, which is around 7 minutes. This prediction time is better than the one given by the previous model to this same group of runners. This estimation error may be high for professional athletes but it is a rather good reference time for amateur runners.

Figure 1 shows hexplots with the results for bagging in the filtered dataset (left plot) and the *injure free* dataset (right plot). Each cell of the plot indicates (with a grey color scale coding) the number of runners with a marathon mark given by the horizontal axis and a prediction given by the vertical axis. It can be observed how the predictions lay along the diagonal and that the largest number of runners reach the finish line between 11000 and 12000 seconds, i.e. between 3:03 and 3:20 hh:mm. In addition, we can observe that runners could be ranked reasonably. However, there is a slight bias in the model in the extremes. The fastest runners tend to get a prediction slower than their real time and a faster prediction is given by the system to the slowest runners.

4 Conclusions

In this paper we have shown how machine learning algorithms can be used for the prediction of marathon performance for amateur athletes. The prediction performance was estimated based on the training programmes of the athletes. Thousands of running trainings were extracted from a sports social network for athletes running the 2015 Boston and London marathons. The training sessions include the distance covered, the time of the training session, the net elevation gain, date and pace. For each runner, the workouts for the 6 weeks period prior to the race were extracted.

Furthermore, in order to be able to feed the data to the machine learning algorithms, a set of uniform attributes had to be extracted for each runner. We splited the training sessions into: six one-week periods, three two-week periods and one six-week period. For each of these periods we extracted the total number of kilometers, the total training time and the total net elevation gain. In addition for the longest and for the fastest training session in each period we extracted the number of kilometers, time, pace and net elevation gain.

The estimation of the performance was analyzed in two sets of experiments. First, we analyzed all the data independently of the number of workouts of each athlete. Despite of the fact that many runners post on their walls only a fraction of their training sessions, a mean absolute error roughly above 11 minutes was obtained using bagging and the whole set of extracted attributes. In a second experiment, we tested the proposed procedure by taking into account only those athletes with a reasonable number of workouts per week (that was set to three workouts per week). The prediction time estimation for this experiment dropped to nine minutes on average. Finally, by training on a dataset using only runners who supposedly did not have any problem during the race, a mean average error of seven minutes is obtained, which is a fairly good performance prediction for amateur runners.

References

1. Run-down running statistics & rankings. http://run-down.com/statistics/, accessed: 2016-07-03
2. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Koetter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B.: Knime. the konstanz information miner. SIGKDD Explorations 11(1) (2009)
3. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
4. Daniels, J.: Daniels' Running Formula. Human Kinetics (2005)
5. Gardner, J.B., Gerry Purdy, J.: Computer generated track scoring tables. Medicine and Science in Sports and Exercise 2(3), 152–161 (1970)
6. Godsey, B.: Comparing and forecasting performances in different events of athletics using a probabilistic model. Journal of Quantitative Analysis in Sports 8 (2012)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11(1) (2009)
8. Scheerder, J., Breedveld, K., Borgers, J.: Running across Europe: The rise and size of one of the largest sport markets. Houndmills, Basingstoke, Hampshire; New York : Palgrave Macmillan (2015)
9. Scheerder, J., Breedveld, K., Borgers, J.: Who is doing a run with the running boom?: The growth and governance of one of Europes most popular sport activities (2015)
10. Tejero-Gonzlez, C.M.: The number of runners in spain increased during the first decade of the 21st century. Apunts. Educacin Fsica y Deportes 120, 73–75 (2015)
11. USA, R.: 2015 running usa annual marathon report. Tech. rep., Running USA (2016), http://www.runningusa.org/marathon-report-2016